Cueing Without Gapping: Cuer-Independent Cued Speech Recognition Powered by Cross-Cuer Invariant Modeling

Fengji Ma[1], **Chenxing Li**[2], **Li Liu**[1]*

[1]Hong Kong University of Science and Technology (Guangzhou)
[2]Tencent AI Lab
avrillliu@hkust-gz.edu.cn

Abstract

Automatic Cued Speech Recognition (ACSR) is a vital communication system designed to enhance spoken language accessibility for the hearing-impaired by combining lip movements and hand gestures to encode phonemes. Despite its effectiveness, current ACSR methods face significant challenges, including poor generalization to unseen cuers[1] due to the limited scale of CS datasets, which restricts the ability of existing visual encoder to capture cuer-invariant CS visual features. Additionally, previous approaches relying on Connectionist Temporal Classification (CTC) decoding fail to incorporate prior linguistic sequence knowledge, further limiting their performance. To address these issues, we propose a novel **T**wo **A**uxiliary Modalities guided **C**ross-cuer **I**nvariant **A**daptation method (**TACIA**), introducing pose and text modalities to help extract cuer-invariant motion and semantic features, thereby improving generalization. In addition, we introduce a Visual-guided Cued Token Prediction (VG-NTP) method, inspired by large language models. This method replaces CTC decoding by incorporating language modeling, leveraging rich linguistic knowledge, including semantics, to address the suboptimal issues present in the CTC decoding process. Extensive experiments demonstrate the superiority of our approach to the state-of-the-art (SOTA) on Chinese and British CS datasets, significantly advancing the accuracy and quality of ACSR systems.

Introduction

Cued Speech (CS) is an efficient communication system designed to enhance the linguistic accessibility of spoken language for the hearing-impaired. Unlike traditional lip-reading, which often suffers from insufficient information, CS employs a combination of lip movements and several hand gestures to encode spoken language, significantly improving communication efficacy. To date, CS has been adapted into over 60 languages, including English, French, and Mandarin, highlighting its growing importance as an effective communication tool for the deaf and hard of hearing.

As shown in Fig. 1, Liu et al. (Liu and Feng 2019) pioneered the first Mandarin CS system, utilizing five hand positions (mouth, chin, throat, side, and cheek) to encode all Chinese vowel groups and eight hand shapes to represent Chinese consonants. It is noted that Sign Language (SL) is another widely used communication method, as noted in various studies (Stokoe 2005; Liddell and Johnson 1989; Timothy 2003; Zhou et al. 2023; Li et al. 2025; Chen et al. 2024). However, it is crucial to clarify that CS is not a visual language like SL; rather, it functions as a coding system for spoken language (Cornett 1967). Furthermore, research has shown that CS can be acquired significantly faster than SL (Reynolds 2007). Additionally, compared to text-based communication, CS is more accessible and easier to adopt for hearing-impaired individuals who may be illiterate (Cox et al. 2002; Power, Power, and Rehling 2007).

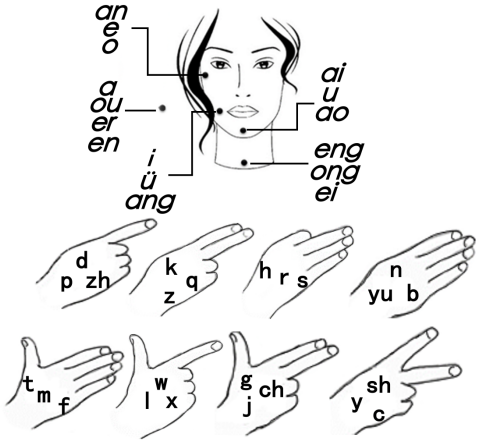
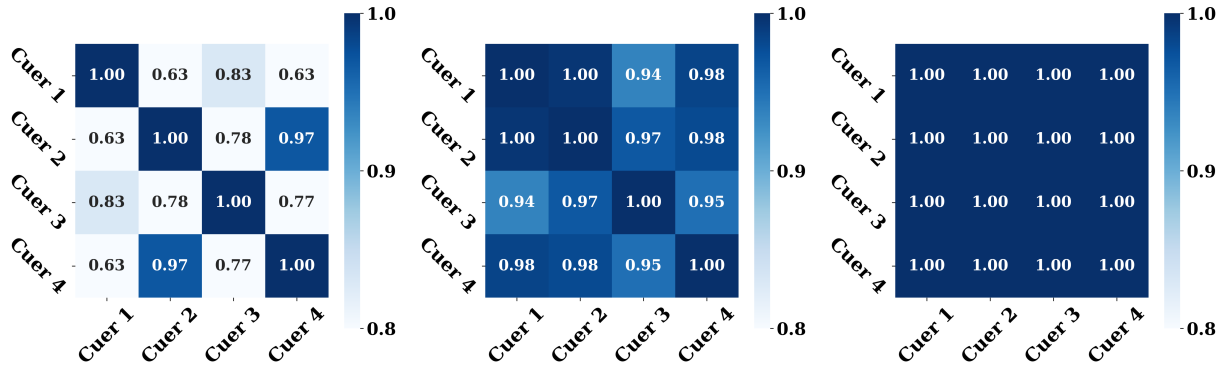


Figure 1: The Mandarin Chinese CS system (from (Liu and Feng 2019)).

In the literature, early research on automatic CS recognition (ACSR) mainly focused on the multimodal feature fusion of hand and lip movements. For example, Heracleous et al. (Heracleous, Beautemps, and Hagita 2012) directly fused lip shape and hand position features and employed a Hidden Markov Model (HMM) (Rabiner and Juang 1986) to predict phonemes. Sankar et al. (Sankar, Beautemps, and Hueber 2022) proposed a multi-stream Convolutional Neural Network (CNN) to extract and fuse features from hand and lip modalities, using an HMM-GMM phoneme decoder to predict phonemes. To address the insufficient capture of

*Corresponding Author

[1]People who perform CS are called cuer.



(a) Visual (b) Pose (c) Text

Figure 2: Similarity matrix between different cuers for the same sentence across different modalities (visual, pose, and text). A comparison of (a) to (c) reveals that for the same sentence, the visual features in (a) vary significantly across different cuers (i.e., cuer-dependent). These differences stem from visual elements such as shape. However, CS pose and semantic text features in our method effectively eliminate this variability (i.e., not cuer-independent).

global dependencies in multimodal data and the computational inefficiency of existing methods, Liu et.al (Liu, Liu, and Li 2024) first model all tokens for each modality to capture modality-specific fine-grained temporal dependencies while learning coarse-grained temporal dependencies shared across modalities. Efficient cross-modal interaction is achieved by selecting important tokens.

Despite these advancements, current methods face two main limitations. **Firstly**, due to the limited scale of CS datasets, existing ACSR methods struggle to generalize to unseen cuers. Different cuers exhibit different cue habits, and there are significant visual feature variations between different cuers, such as shape differences across cues (as shown in Fig. 2a), which further restrict the ability of the visual encoder to extract cuer-invariant features. **Secondly**, the commonly used Connectionist Temporal Classification (CTC) decoding in ACSR (Liu and Liu 2023; Liu, Liu, and Li 2024) often leads to suboptimal results. Researches (Komatsu et al. 2022; Higuchi et al. 2022) found that CTC lacks explicit language modeling capabilities and does not directly leverage rich linguistic knowledge, including semantics. Previous ACSR work mostly focused on CTC decoding and did not incorporate language modeling in ACSR decoding stage.

To address the above two challenges, we propose a novel **T**wo **A**uxiliary Modalities guided **C**ross-cuer **I**nvariant **A**daptation method (**TACIA**). The key assumption behind this is that for the same word or sentence, when prompted by different individuals, the standardized CS motions should exhibit consistent pose characteristics. By leveraging pose and text auxiliary modalities, this method enables the visual encoder to effectively extract cuer-invariant features while reducing excessive reliance on the cuer's cueing habits. When processing the pose data, we further account for inaccuracies or motion distortions caused by rapid hand movements and the resulting motion blur or ghosting artifacts, which often lead to missing or inaccurate pose estimates in the video. This approach helps mitigate hand distortion problems. Furthermore, inspired by large language models (LLMs), we replace the CTC decoding mechanism with a Visual-Guided Next cued Token Prediction (VG-NTP) process. This innovative decoding strategy leverages next token prediction to better align video sequences with textual output, overcoming the limitations of CTC and improving the overall performance of the ACSR system.

In summary, the main contributions of this work are as follows:

- To tackle the generalization issue of unseen cuers in ACSR, we design a multimodal cuer-invariant adaptation method, which introduces pose and text as auxiliary modalities to help the visual encoder effectively extract visual features that are independent of the different cuers.
- We introduce a novel decoding strategy, VG-NTP, to replace the traditional CTC decoding mechanism. Based on LLMs, the VG-NTP predicts the next token to better align video sequences with textual outputs, improving the overall performance of the ACSR. To the best of our knowledge, this is the first work that introduces the LM knowledge to ACSR decoding.
- Extensive experiments validate the superiority of our method to previous state-of-the-art (SOTA) in both accuracy and sentence generation quality. Specifically, in the Chinese CS dataset, our approach achieves a remarkable 16.6% reduction in Word Error Rate (WER) and a 4.2 improvement in BLEU score.

Related Works

Automatic Cued Speech Recognition

Existing research of ACSR primarily focuses on multimodal fusion, especially extracting and connecting multimodal features. For instance, studies (Heracleous, Beautemps, and Hagita 2012; Wang et al. 2021a) marked regions of interest (ROI) on the lips and hands to extract visual features, which were then directly fused for cross-modal align-

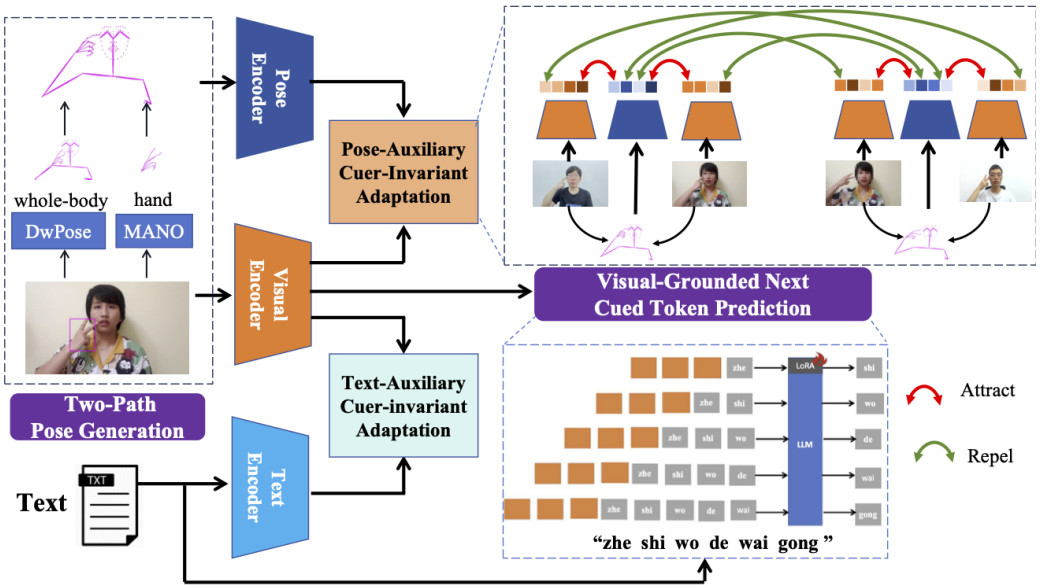


Figure 3: Framework of our proposed method (TACIA). It incorporates a two-path pose generation strategy, utilizing encoders for both hand and whole-body poses to generate accurate and invariant pose representations. The pose encoder extracts pose features, which undergo a pose-auxiliary cue-invariant adaptation process to ensure that the same gestures across different cuers are aligned. The visual encoder processes image frames to capture the visual features, which are then linked to the text tokens through a VG-NTP.

ment. In (Gao, Huang, and Liu 2023), a multimodal alignment resynchronization process was proposed, requiring statistical pre-definition of hand precedence timing in a CS dataset. Unlike existing ACSR approaches that require preprocessing of the lip and hand ROIs, the method proposed in this paper does not necessitate such pre-processing for lip and hand ROIs, greatly simplifying the preprocessing steps. Furthermore, to address the unseen cuer generalization problem, as motivated by Fig. 2, we propose a cue-invariant adaptation method utilizing both pose and text modalities, aligning visual-pose and visual-text information to extract as much motion and semantic information as possible. To the best of our knowledge, this is the first work to introduce a pose modality into ACSR. Additionally, we address issues such as distortion, deformation, and pose-estimation errors arising from fast motion blur. This approach effectively mitigates these challenges, ensuring more accurate pose representation for ACSR tasks.

Lip Reading and Sign Language Recognition

Next, we will introduce recent advancements in Lip Reading and SL recognition, which are closely related to CS.

Lip reading tasks face challenges in generalizing to unseen cuers. To address this, landmarks have been employed to capture fine-grained speaker-invariant visual cues. (Wu et al. 2024) proposed a landmark-guided lip reading method with max-min mutual information regularization to enhance generalization. (Xue et al. 2023) introduced the LipFormer, utilizing a visual-landmark Transformer to fuse lip movements with facial features, reducing speaker appearance biases. (Qu, Weber, and Wermter 2022) developed LipSound2, an encoder-decoder model with position-aware attention that maps facial sequences to speech, achieving high-quality lip reading through self-supervised pretraining. However, these methods focus only on facial and lip landmarks, neglecting hand gestures and positions. In contrast, in ACSR, we propose integrating skeleton poses, encompassing both facial and hand features, to better represent visual dynamics.

SL, as a visual language, bridges motions and semantics using glosses. (Fayyazsanavi, Anastasopoulos, and Kosecka 2024) enhanced gloss-to-text translation by combining pretrained language models, data augmentation, and semantically aware loss. (Moryossef et al. 2023) developed a gloss-based text-to-sign benchmark. (Guo et al. 2024) improved SL recognition with gloss-prior-guided visual features, while (Camgoz et al. 2020) used a Transformer for end-to-end SL recognition and translation. However, unlike lip reading and CS, where sentence expression relies on the order of words, SL follows a special spatial-visual language encoding system. It depends on the body position, direction, movement of gestures, as well as facial expressions, and can clarify the subject and object of motion through the direction and position of gestures without relying on a fixed word order. At the same time, SL does not depend on lip reading, and these differences make it difficult to directly apply SL recognition methods to our ACSR.

Methodology

Feature Extraction

Text Encoder. Our text encoder module, ψ, is built upon the mBART architecture (Liu et al. 2020) to convert tokenized

inputs (both English and Pinyin) into their corresponding semantic embeddings. This encoder supports both English and Pinyin-based Mandarin. We keep the parameters of this encoder frozen which ensures that the model leverages the pretrained linguistic priors from mBART without modification. For Mandarin cued speech, we fine-tune only the tokenizer on Pinyin transcriptions, keeping the network weights unchanged.

Visual Encoder. The visual encoder ϕ comprises a frame-level backbone and a lightweight temporal modeling module. Spatial features for each video frame are processed using a ResNet backbone (He et al. 2016a), which was pretrained on ImageNet (Deng et al. 2009). Local temporal dynamics are subsequently captured by a lightweight temporal module. This module comprises two stacked blocks, each applying temporal convolution followed by batch normalization and ReLU activation. This design enables the encoder to capture local dynamics without compromising computational efficiency.

Two-Path Pose Generation. To extract upper-body motion representations, we apply DWPose (Yang et al. 2023) to detect 2D landmarks based on the COCO-WholeBody format (Jin et al. 2020). We retain only the facial and hand keypoints, omitting the lower body in order to better cope with missing or inaccurate keypoints and motion blur. We additionally incorporate MANO-based hand estimation (Moon 2023). The resulting keypoints are processed into skeletal heatmaps, serving as pose input. The entire process is a two-path pose generation module, denoted f_{pose}, which provides separate face and hand streams.

Pose Encoder. The pose encoder ϕ_p is designed to extract cuer-invariant motion features from facial and hand sequences. Given a series of keypoints from each video, we decompose the sequence into two parallel branches. The hand stream is processed using a spatial-temporal graph convolutional network (ST-GCN) (Yan, Xiong, and Lin 2018), which models motion through a fixed skeletal graph structure. The facial stream is handled by a stack of temporal 1D convolutions, enabling the encoder to capture temporal variations in lip and facial expression. Both streams are globally pooled and concatenated, followed by a linear projection into the shared latent space. This fused representation is used for subsequent multi-modal alignment.

Two-Modality Cuers-Invariant Adaptation

To address the challenge of visual variations across different cuers, we introduce a **Two-Modality Cuer-Invariant Adaptation** strategy. The model is guided by two auxiliary modalities, **text** and **pose**, both of which tend to remain more stable across cuers, thereby enhancing the learning process. For the same utterance, its semantic content (represented by text) and its standardized motion patterns (represented by pose) remain constant. By aligning the primary visual features with these stable auxiliary features through contrastive learning, our model learns to extract visual representations that are robust to cuer-specific habits and appearances.

Text-Auxiliary Cuer-Invariant Adaptation. This module leverages the semantic content of the accompanying text to guide the learning of cuer-invariant visual features. The underlying principle is that for a given sentence, its semantic meaning is constant regardless of which cuer performs it. Therefore, the visual representation of a cued sentence should be closely aligned with its corresponding text representation.

Specifically, we treat as positive pairs the visual embeddings from different cuers performing the same sentence and the corresponding text embeddings. Visual and text embeddings from different sentences are treated as negative pairs. Let $T^{(i)}$ denote the tokenized text sequence of the i-th sentence, and $V^{(i,c)}$ denote the visual sequence of the same sentence performed by cuer c. The projected embeddings are computed as $z_{\text{text}}^{(i)} = f_{\text{text}}(\psi(T^{(i)}))$, and $z_{\text{visual}}^{(i,c)} = f_{\text{visual}}(\text{AvgPool}(\phi(V^{(i,c)})))$, where $\psi(\cdot)$ and $\phi(\cdot)$ are the text and visual encoders respectively, and f_{text}, f_{visual} are projection functions. We define the set of **positive pairs** $\mathcal{S}_+^{\text{text}}$ and **negative pairs** $\mathcal{S}_-^{\text{text}}$ as:

$$\begin{aligned}\mathcal{S}_+^{\text{text}} &= \left\{ \left(z_{\text{text}}^{(i)}, z_{\text{visual}}^{(i,c)} \right) \mid \forall i, c \right\}, \\ \mathcal{S}_-^{\text{text}} &= \left\{ \left(z_{\text{text}}^{(i)}, z_{\text{visual}}^{(j,c')} \right) \mid i \neq j,\ \forall c' \right\}. \end{aligned} \tag{1}$$

The InfoNCE-based contrastive loss is defined as:

$$\mathcal{L}_{\text{text}} = -\frac{1}{2|\mathcal{S}_{\text{text}}^+|} \sum_{(i,c)\in\mathcal{S}_{\text{text}}^+} \left(\log p_{\text{visual}\to\text{text}}^{(i,c)} + \log p_{\text{text}\to\text{visual}}^{(i,c)} \right), \tag{2}$$

where $p_{\text{visual}\to\text{text}}^{(i,c)}$ and $p_{\text{text}\to\text{visual}}^{(i,c)}$ are as follows:

$$\begin{aligned} p_{\text{visual}\to\text{text}}^{(i,c)} &= \frac{\exp(z_{\text{visual}}^{(i,c)} \cdot z_{\text{text}}^{(i)}/\tau)}{\sum_{j=1}^{\mathcal{B}} \exp(z_{\text{visual}}^{(i,c)} \cdot z_{\text{text}}^{(j)}/\tau)}, \\ p_{\text{text}\to\text{visual}}^{(i,c)} &= \frac{\exp(z_{\text{visual}}^{(i,c)} \cdot z_{\text{text}}^{(i)}/\tau)}{\sum_{j,c'} \exp(z_{\text{visual}}^{(j,c')} \cdot z_{\text{text}}^{(i)}/\tau)}. \end{aligned} \tag{3}$$

Pose-Auxiliary Cuer-Invariant Adaptation. Similarly, this module utilizes skeletal pose information to achieve cuer-invariance. The key assumption is that the standardized CS system dictates a consistent motion pattern (i.e., hand shape, position, and trajectory) for a given phoneme sequence. While a cuer's physical appearance or subtle stylistic habits might vary, the underlying skeletal motion should remain consistent. This module aligns the embeddings of visual and pose sequences in a shared latent space, ensuring invariance across different cuers. Positive pairs consist of visual embeddings and pose embeddings from different cuers performing the same sentence. Negative pairs are formed from sequences corresponding to different sentences. Let $V^{(i,c)}$ denote the visual sequence of sentence i performed by **cuer** c, and $P^{(i,c')}$ denote the pose sequence of the same sentence performed by **cuer** c'. The projected embeddings are defined as $z_{\text{visual}}^{(i,c)} = f_{\text{visual}}(\text{AvgPool}(\phi(V^{(i,c)})))$ and $z_{\text{pose}}^{(i,c')} = f_{\text{pose}}(\text{AvgPool}(\phi_p(P^{(i,c')})))$, where $\phi(\cdot)$ and $\phi_p(\cdot)$ are the visual and pose encoders respectively. The set of **positive pairs** $\mathcal{S}_+^{\text{pose}}$ and **negative pairs** $\mathcal{S}_-^{\text{pose}}$ are:

$$\begin{aligned}\mathcal{S}_+^{\text{pose}} &= \left\{ \left(z_{\text{visual}}^{(i,c)}, z_{\text{pose}}^{(i,c')} \right) \mid c \neq c' \right\}, \\ \mathcal{S}_-^{\text{pose}} &= \left\{ \left(z_{\text{visual}}^{(i,c)}, z_{\text{pose}}^{(j,c')} \right) \mid i \neq j \right\}. \end{aligned} \tag{4}$$

The InfoNCE-based contrastive loss is given by:

$$\mathcal{L}_{\text{pose}} = -\frac{1}{2|\mathcal{S}^{+}_{\text{pose}}|} \sum_{(i,c,c') \in \mathcal{S}^{+}_{\text{pose}}} \left(\log p^{(i,c,c')}_{\text{visual}\to\text{pose}} + \log p^{(i,c,c')}_{\text{pose}\to\text{visual}} \right), \quad (5)$$

where $p^{(i,c,c')}_{\text{visual}\to\text{pose}}$ and $p^{(i,c,c')}_{\text{pose}\to\text{visual}}$ are as follows:

$$\begin{aligned} p^{(i,c,c')}_{\text{visual}\to\text{pose}} &= \frac{\exp(z^{(i,c)}_{\text{visual}} \cdot z^{(i,c')}_{\text{pose}}/\tau)}{\sum_{j,\tilde{c}} \exp(z^{(i,c)}_{\text{visual}} \cdot z^{(j,\tilde{c})}_{\text{pose}}/\tau)}, \\ p^{(i,c,c')}_{\text{pose}\to\text{visual}} &= \frac{\exp(z^{(i,c)}_{\text{visual}} \cdot z^{(i,c')}_{\text{pose}}/\tau)}{\sum_{j,\tilde{c}} \exp(z^{(j,\tilde{c})}_{\text{visual}} \cdot z^{(i,c')}_{\text{pose}}/\tau)}. \end{aligned} \quad (6)$$

Visual-Guided Next Cued Token Prediction

We built a visual encoder followed by a transformer decoder to perform CS recognition through a video-based text generation task. In CS videos, the video frame number is substantially higher than the corresponding token or character rate, leading to significant data redundancy and causing redundant visual evidence per token. To improve computational efficiency, we uniformly downsampled the video sequence to bound latency.

To model the short-term dynamics of gestures and lip movements, the frame-level features from the visual backbone are passed through a temporal module. This module comprises stacked blocks of 1D temporal convolution, followed by batch normalization and ReLU activation, which is utilized to aggregate sub-token dynamics into token-synchronous features, enhancing module's context. For ease of representation, we denote the pre-trained visual encoder and temporal module together as the visual encoder. The above operations can be expressed as $h_{1:N} = \phi(V_{1:N})$, where ϕ represents the video encoder (visual encoder). The decoder conditions on visual states $h_{1:N}$ and previously generated cued tokens to autoregressively predict the next token, *i.e.*, $w_i = \Psi(w_{1:i-1}, h_{1:N})$, aligning with CS-specific visual evidence. Finally, the output features from the decoder are passed through a linear projection layer (the language model head W, b) and a softmax function f_{sm} to compute the probability distribution as follows:

$$p(o_i|o_{1:i-1}, V) = f_{sm}(W \cdot w_i + b), \quad (7)$$

where o_i represents the output token. The training objective for the decoding module is to maximize the likelihood of the ground-truth text sequence given the input video. This is achieved by optimizing a standard sequence-to-sequence cross-entropy loss between the model's predictions and the target sentences as follows:

$$\mathcal{L} = -\sum_{i=1}^{L} \log p(o_i|o_{1:i-1}, V). \quad (8)$$

We utilized Low-Rank Adaptation (LORA) (Hu et al. 2022) for parameter-efficient fine-tuning of the language model. This technique introduces a small set of trainable, low-rank matrices into the model's layers while keeping the original pre-trained weights frozen. A primary advantage is the preservation of the LLM's foundational knowledge and capabilities during adaptation.

Experiments

In this section, we first provide details on the experimental setup and benchmark datasets. Then, we thoroughly evaluate the impact of our proposed method, including the ablation study, comparing our results with SOTA methods across two benchmark datasets.

Datasets and Metrics

Datasets. We evaluate our proposed method on two mainstream datasets: We conducted experiments on two publicly available CS datasets, namely the Mandarin Chinese (Liu and Feng 2019) and British English (Sankar, Beautemps, and Hueber 2022) CS datasets. The Mandarin Chinese CS dataset is the first large-scale multi-cuer CS dataset for Mandarin Chinese. It contains 6,000 sentences from 6 cuers, where each cuer contributed 1,000 sentences. Chinese phonemes are represented using 40 phonemes (including vowels and consonants), 8 hand shapes, and 5 hand positions alongside lip movements. The British English dataset comprises 390 sentences from 5 cuers. Of these, 43 sentences are shared among all cuers, while the remaining 57 are from a subset of cuers. British English phonemes are represented by 44 phonemes, 8 hand shapes, and 4 hand positions, along with corresponding lip movements.

Evaluation Metrics. Following previous works (Liu and Liu 2023; Liu, Liu, and Li 2024), we adopt the Word Error Rate (WER), which measures the percentage of incorrect words in the recognized text, and the Character Error Rate (CER) for the recognition task. Besides error rate values, we choose BLEU (Papineni et al. 2002), which assesses the quality of text based on n-gram overlap and longest common subsequences. Lower CER and WER indicate more accurate recognition results, while higher BLEU signifies better text quality. The metric that measures n-gram rules and performs sentence quality evaluation is denoted as BLEU@n.

Experimental Setup

Implementation Details. Our visual backbone utilizes a ResNet architecture (He et al. 2016a), which was pre-trained on the ImageNet dataset (Deng et al. 2009) to extract spatial features from each frame. Following the ResNet backbone, we discarded the final fully connected layer and appended a temporal modeling module. This module consists of 1D convolutions configured with a kernel size of 3 and a stride of 1 to capture local temporal dynamics. For this pretraining stage, our transformer architecture included both an encoder and a decoder, each constructed with 3 layers. Within each layer, we configured 8 attention heads, a 512-dimension hidden state, and a feed-forward network dimension of 2048. A dropout rate of 0.1 was applied throughout the transformer blocks to mitigate overfitting. We trained the VG-NTP module using SGD with a 0.9 momentum value. The learning rate was managed by a cosine annealing scheduler, starting at an initial value of 1×10^{-2} and gradually decaying to 0 over the training epochs.

Methods	Chinese		British
	WER ↓	CER ↓	CER ↓
CNN+LSTM (Papadimitriou and Potamianos 2021)	96.3	86.6	94.7
CNN+CTC (He et al. 2016b)	94.4	81.8	94.1
JLF + COS + CTC (Wang et al. 2021b)	91.3	82.2	78.3
Self-attention (Vaswani 2017)	92.2	83.4	79.2
CMML (Liu and Liu 2023)	73.9	58.1	75.1
EcoCued (Liu, Liu, and Li 2024)	71.5	54.4	73.6
TACIA (Ours)	**57.3**	**34.6**	**67.2**

Table 1: Quantitative evaluation of ACSR task in both Chinese and British English CS datasets. CER and WER are adopted as the evaluation metrics. WER is omitted for the British CS Dataset due to missing word-level timestamps in the public release. Following prior work (Liu and Liu 2023), we report CER as a reliable sentence-level metric. The results highlight our method's ability to generate accurate and coherent sentences.

Methods	BLEU@1↑	BLEU@2↑	BLEU@3↑	BLEU@4↑
CMML (Liu and Liu 2023)	28.73	19.77	14.83	10.98
EcoCued (Liu, Liu, and Li 2024)	33.69	24.03	16.91	12.03
TACIA (Ours)	**40.73**	**29.33**	**21.06**	**15.18**

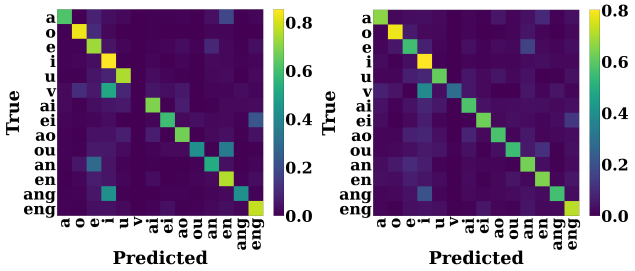
Table 2: BLEU results on Mandarin Chinese CS dataset. We bold the highest scores.

Comparison with SOTA Methods

From Table. 1 and Table. 2, our method outperforms CMML and EcoCued on the Mandarin Chinese/British CS dataset, achieving the lowest WER and CER. In terms of BLEU scores, our method achieves a BLEU@4 of 15.18, surpassing 12.03 in EcoCued (Liu, Liu, and Li 2024) and 10.98 in CMML (Liu and Liu 2023). Additionally, at BLEU@1, our score of 40.73 exceeds EcoCued's 33.69 and CMML's 28.73. For Mandarin Chinese CS, our method achieves a WER of 57.3% and CER of 34.6%, outperforming CMML (Liu and Liu 2023) (73.9% WER, 58.1% CER) and EcoCued (Liu, Liu, and Li 2024) (71.5% WER, 54.4% CER). For British English CS, only CER is reported due to missing word-level timestamps. Our method obtains a CER of 67.2%, which is better than CMML (75.1%) and EcoCued (73.6%), indicating improved generalization across cuers. These results emphasize our method's strong performance in generating accurate and fluent sentences, with improvements in both linguistic accuracy and contextual coherence.

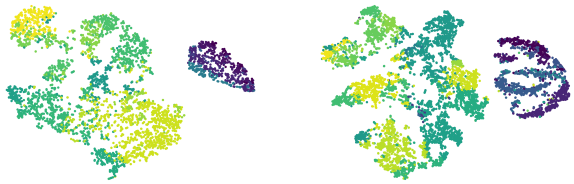
Visualization Analysis

Confusion Matrix Analysis. The confusion matrices presented in Fig. 4a and Fig. 4b compare the performance of CMML and our proposed TACIA method on unseen cuers, focusing on the classification of Mandarin Cued Speech (CS) consonants and vowels. In Fig. 4a, while the CMML baseline establishes a general diagonal pattern, there is noticeable off-diagonal noise. Specifically, the model exhibits significant confusion between phonemes such as "en" and "a", as well as "ang" and "i", suggesting that the acoustic-visual features captured by CMML are insufficient to fully distinguish between these subtle phonetic variations.



(a) CMML (b) TACIA (Ours)

Figure 4: Confusion matrix on unseen cuers with different methods.



(a) CMML (b) TACIA (Ours)

Figure 5: Comparative t-SNE visualization of visual embeddings for unseen cuers.

In contrast, Fig. 4b demonstrates the superior performance of our TACIA method. The diagonal clarity is markedly higher, with most phonemes—such as "o", "e", "i", and "ei". The misclassifications present in the baseline are substantially suppressed, indicating that our method effectively leverages the multi-modal information to form more robust decision boundaries for finals. Even for traditionally difficult pairs like "en" and "eng", TACIA maintains high diagonal purity, reflecting a more precise extraction of linguistic features from the combined lip and hand modalities.

tSNE Visualization. Fig. 5 visualizes the t-SNE embeddings of character representations for unseen cuers, comparing the baseline CMML with our proposed TACIA. This visualization provides intuitive insight into the model's ability to generalize and maintain feature discriminability on novel speakers. As shown in Fig. 5 (a) (CMML), the feature distribution exhibits significant dispersion and entanglement. The decision boundaries between different character categories are blurred, particularly in the central region where clusters overlap substantially. This suggests that CMML struggles to disentangle character identity from cuer-specific variations, leading to "noisy" representations that hinder performance on unseen cuers. In contrast, our TACIA method in Fig. 5 (b) demonstrates superior clustering quality. The character embeddings form highly compact and dense clusters with sharp distinct margins. The clear separation between categories—evident from the isolated clusters and the structured distribution—indicates that TACIA effectively extracts cuer-invariant linguistic features. By minimizing the interference of individual cuer characteristics, our method main-

Auxiliary		WER↓	CER↓	BLEU@4↑
Text	Pose			
✗	✗	67.1	37.8	12.07
✗	✓	65.5	36.1	13.25
✓	✗	58.7	35.9	13.49
✓	✓	**57.3**	**34.6**	**15.18**

Table 3: Ablation study of auxiliary modalities.

tains robust discriminability between characters (e.g., distinct phoneme boundaries) even when processing data from unseen subjects, thereby achieving substantially better generalization.

Ablation Study

Effect of Auxiliary and Input Modalities. The ablation study in Table 3 highlights the critical role of auxiliary and input modalities in reducing WER. Using only visual input yields a baseline WER of 67.1%. Adding pose as an auxiliary modality reduces WER to 65.5%, showing that motion cues effectively complement visual input. Further integrating text and pose auxiliary modalities with visual input lowers WER to 58.7%, as text provides additional linguistic context. Combining both text and pose auxiliaries with visual input achieves the best WER of 57.3% and the highest BLEU@4, indicating that semantic and motion cues are complementary. Overall, combining visual input with text and pose auxiliary modalities achieves optimal recognition performance.

Effect of Language Modeling and Computational Demands. Tab. 4 presents a comprehensive ablation study evaluating the impact of incorporating LLMs into the VG-NTP framework. The table reports the number of parameters, inference latency, and the performance in terms of WER and BLEU@4. Without any language model ("w/o LLM"), the system relies solely on CTC decoding, resulting in a WER of 68.5% and a BLEU@4 score of 11.54. Introducing mT5-Small (300M parameters) already brings noticeable improvements, reducing the WER to 61.2% and increasing BLEU@4 to 13.8, with only a moderate increase in inference time (0.85s). As the model size increases, performance continues to improve: mBART-Large (610M) achieves a WER of 57.3% and BLEU@4 of 15.18, while mT5-Large (1.2B) further reduces the WER to 54.6% and raises BLEU@4 to 19.32. The best performance is observed with mT5-XL (3.7B), which achieves a WER of 51.5% and a BLEU@4 score of 22.5. However, this comes at the cost of significantly higher inference time (3.20s), indicating a clear trade-off between accuracy and computational efficiency. The results demonstrate a consistent trend that larger language models lead to better recognition performance. Nonetheless, increasing model size also brings additional inference overhead. Thus, in practical deployments, one must balance performance gains with latency constraints. These findings highlight the potential of scaling LLMs for improved ACSR accuracy, while also motivating future exploration of more efficient model architectures or distillation techniques to reduce latency. However, due to computational resource constraints, we were unable to experiment with even larger models, such as LLAMA-7B (Touvron et al. 2023). Nevertheless, the experimental results in the table underscore the promising potential of applying larger parameter models for further performance enhancement.

LLM	# Parameters	Inference Time (s) ↓	WER (%) ↓	BLEU@4 ↑
w/o LLM	-	0.35	68.5	11.54
mT5-Small	300M	0.85	61.2	13.80
mBART-Large	610M	1.14	57.3	15.18
mT5-Large	1.2B	1.65	54.6	19.32
mT5-XL	3.7B	3.20	51.5	22.50

Table 4: Comparison of different LLMs in terms of inference time per frame, WER, and BLEU@4.

Method	Pose Masking Ratio	WER (%) ↓	BLEU@4 ↑
CMML (Baseline)	-	73.9	10.98
TACIA (Ours)	0%	57.3	15.18
TACIA (Ours)	30%	57.9	14.95

Table 5: Impact of pose masking ratio.

Analysis of Pose Robustness. We further investigate robustness to missing pose information by training TACIA with randomly masked pose keypoints, where a fixed percentage of keypoints is removed throughout training. As reported in Tab. 5, although performance gradually decreases as the masking ratio increases, our method still clearly outperforms the CMML baseline even when 30% of the pose data are removed during training. This confirms that TACIA can learn effectively from incomplete pose inputs and is robust to pose estimation failures.

Conclusion

This work proposes a novel framework for ACSR, addressing key challenges in generalization and decoding accuracy. By introducing a multimodal adaptation strategy leveraging pose and text modalities, the model effectively extracts cue-invariant features and reduces dependency on cuer-specific attributes. A two-path pose generation method mitigates motion-estimation errors introduced by fast hand movements and motion blur, thereby enhancing robustness. Furthermore, the VG-NTP mechanism replaces traditional CTC decoding, leveraging linguistic context to improve recognition and sentence generation. Extensive experiments demonstrate significant performance improvements, including a 16.6% reduction in WER and a 4.2 point BLEU increase, surpassing SOTA methods. This work highlights the potential of multimodal integration and advanced decoding strategies for improving ACSR systems, paving the way for enhanced accessibility for the hearing impaired.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62471420), Guang-Dong Basic and Applied Basic Research Foundation (2025A1515012296), and 2025 Tencent AI Lab Rhino-Bird Program.

References

Camgoz, N. C.; Koller, O.; Hadfield, S.; and Bowden, R. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *CVPR*, 10023–10033.

Chen, Z.; Zhou, B.; Li, J.; Wan, J.; Lei, Z.; Jiang, N.; Lu, Q.; and Zhao, G. 2024. Factorized Learning Assisted with Large Language Model for Gloss-free Sign Language Translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, 7071–7081.

Cornett, R. O. 1967. Cued speech. *American annals of the deaf*, 3–13.

Cox, S.; Lincoln, M.; Tryggvason, J.; Nakisa, M.; Wells, M.; Tutt, M.; and Abbott, S. 2002. Tessa, a system to aid communication with deaf people. In *Proceedings of the fifth international ACM conference on Assistive technologies*, 205–212.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255. Ieee.

Fayyazsanavi, P.; Anastasopoulos, A.; and Kosecka, J. 2024. Gloss2Text: Sign Language Gloss translation using LLMs and Semantically Aware Label Smoothing. In *EMNLP*, 16162–16171.

Gao, L.; Huang, S.; and Liu, L. 2023. A novel interpretable and generalizable re-synchronization model for cued speech based on a multi-cuer corpus. *Interspeech*.

Guo, L.; Xue, W.; Liu, B.; Zhang, K.; Yuan, T.; and Metaxas, D. 2024. Gloss Prior Guided Visual Feature Learning for Continuous Sign Language Recognition. *IEEE TIP*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016a. Deep residual learning for image recognition. In *CVPR*, 770–778.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016b. Deep residual learning for image recognition. In *CVPR*, 770–778.

Heracleous, P.; Beautemps, D.; and Hagita, N. 2012. Continuous phoneme recognition in cued speech for french. In *EUSIPCO*, 2090–2093. IEEE.

Higuchi, Y.; Yan, B.; Arora, S.; Ogawa, T.; Kobayashi, T.; and Watanabe, S. 2022. BERT Meets CTC: New Formulation of End-to-End Speech Recognition with Pre-trained Masked Language Model. In *EMNLP*.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*.

Jin, S.; Xu, L.; Xu, J.; Wang, C.; Liu, W.; Qian, C.; Ouyang, W.; and Luo, P. 2020. Whole-body human pose estimation in the wild. In *ECCV*, 196–214. Springer.

Komatsu, T.; Fujita, Y.; Lee, J.; Lee, L.; Watanabe, S.; and Kida, Y. 2022. Better intermediates improve CTC inference. *arXiv preprint arXiv:2204.00176*.

Li, Z.; Zhou, W.; Zhao, W.; Wu, K.; Hu, H.; and Li, H. 2025. Uni-Sign: Toward Unified Sign Language Understanding at Scale. In *The Thirteenth International Conference on Learning Representations*.

Liddell, K. S.; and Johnson, E. R. 1989. AMERICAN SIGN LANGUAGE: THE PHONOLOGICAL BASE. *Sign Language Studies*, 195–278.

Liu, L.; and Feng, G. 2019. A pilot study on mandarin chinese cued speech. *American Annals of the Deaf*, 164(4): 496–518.

Liu, L.; and Liu, L. 2023. Cross-modal mutual learning for cued speech recognition. In *ICASSP*, 1–5. IEEE.

Liu, L.; Liu, L.; and Li, H. 2024. Computation and parameter efficient multi-modal fusion transformer for cued speech recognition. *TASLP*.

Liu, Y.; Gu, J.; Goyal, N.; Li, X.; Edunov, S.; Ghazvininejad, M.; Lewis, M.; and Zettlemoyer, L. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8: 726–742.

Moon, G. 2023. Bringing inputs to shared domains for 3D interacting hands recovery in the wild. In *CVPR*, 17028–17037.

Moryossef, A.; Müller, M.; Göhring, A.; Jiang, Z.; Goldberg, Y.; and Ebling, S. 2023. An open-source gloss-based baseline for spoken to signed language translation. In *Proceedings of the Second International Workshop on Automatic Translation for Signed and Spoken Languages*, 22–33.

Papadimitriou, K.; and Potamianos, G. 2021. A fully convolutional sequence learning approach for cued speech recognition from videos. In *EUSIPCO*, 326–330. IEEE.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 311–318.

Power, D. J.; Power, M. R.; and Rehling, B. 2007. German deaf people using text communication: Short message service, TTY, relay services, fax, and e-mail. *American Annals of the Deaf*, 152(3): 291–301.

Qu, L.; Weber, C.; and Wermter, S. 2022. Lipsound2: Self-supervised pre-training for lip-to-speech reconstruction and lip reading. *IEEE TNNLS*, 35(2): 2772–2782.

Rabiner, L. R.; and Juang, B.-H. 1986. An introduction to hidden Markov models. *IEEE ASSP Magazine*.

Reynolds, S. E. 2007. An examination of Cued Speech as a tool for language, literacy, and bilingualism for children who are deaf or hard of hearing.

Sankar, S.; Beautemps, D.; and Hueber, T. 2022. Multi-stream neural architectures for cued speech recognition using a pre-trained visual feature extractor and constrained ctc decoding. In *ICASSP*, 8477–8481. IEEE.

Stokoe, J., C. William. 2005. Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf. *The Journal of Deaf Studies and Deaf Education*, 10(1): 3–37.

Timothy, R. 2003. LINGUISTICS OF AMERICAN SIGN LANGUAGE: AN INTRODUCTION. *Studies in Second Language Acquisition*, 25(1): 157–158.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.;

Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv*.

Vaswani, A. 2017. Attention is all you need. *NeurIPS*.

Wang, J.; Gu, N.; Yu, M.; Li, X.; Fang, Q.; and Liu, L. 2021a. An Attention Self-Supervised Contrastive Learning Based Three-Stage Model for Hand Shape Feature Representation in Cued Speech. In *Interspeech 2021*.

Wang, J.; yue Tang, Z.; Li, X.; Yu, M.; Fang, Q.; and Liu, L. 2021b. Cross-Modal Knowledge Distillation Method for Automatic Cued Speech Recognition. In *Interspeech*.

Wu, L.; Zhang, X.; Zhang, Y.; Zheng, C.; Liu, T.; Xie, L.; Yan, Y.; and Yin, E. 2024. Landmark-Guided Cross-Speaker Lip Reading with Mutual Information Regularization. In *LREC-COLING*, 10023–10033.

Xue, F.; Li, Y.; Liu, D.; Xie, Y.; Wu, L.; and Hong, R. 2023. Lipformer: learning to lipread unseen speakers based on visual-landmark transformers. *IEEE TCSVT*, 33(9): 4507–4517.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, volume 32.

Yang, Z.; Zeng, A.; Yuan, C.; and Li, Y. 2023. Effective whole-body pose estimation with two-stages distillation. In *ICCV*, 4210–4220.

Zhou, B.; Chen, Z.; Clapés, A.; Wan, J.; Liang, Y.; Escalera, S.; Lei, Z.; and Zhang, D. 2023. Gloss-free sign language translation: Improving from visual-language pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20871–20881.